Generative AI: Beyond the Hype

A progress review, emerging use cases, and a prediction

todd underwood - toddunder@gmail.com

2024-10-30 - SRECon EU 2024 - Dublin, Ireland

Prelude #1:

Who am I?

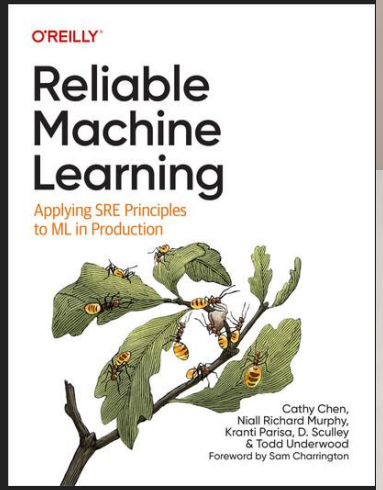
Who are you?[*]

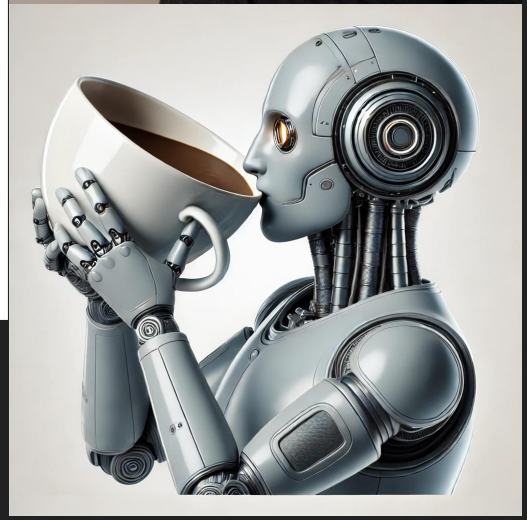
Why are we all here?[**]

[**] Who do I imagine you to be
* I’m obviously not going to really answer that one. You knew that already.
Also, why am I never leaving well enough alone?

Context & Disclaimers: Me

- Worked on large ML training and serving systems at Google and OpenAI for the past 15 years.
- Wrote a book on this stuff (with a bunch of other smarter people).
- AI sceptic (but less than some).
- Not sure you should take me seriously.





Also, Context and Disclaimers: You

- Technical audience who build and run big systems, some of them using ML.
- Annoyed at incessant AI hype (alongside failure to deliver)
- Amused (impressed?) by AI fun use cases. Possibly curious about uses.
- Want to understand the gap (promise-reality) and whether it will ever close.

Disclaimer: not really you. I hope.
All images generated with DALL-E unless otherwise noted

Prelude #2:

Terms and Scope

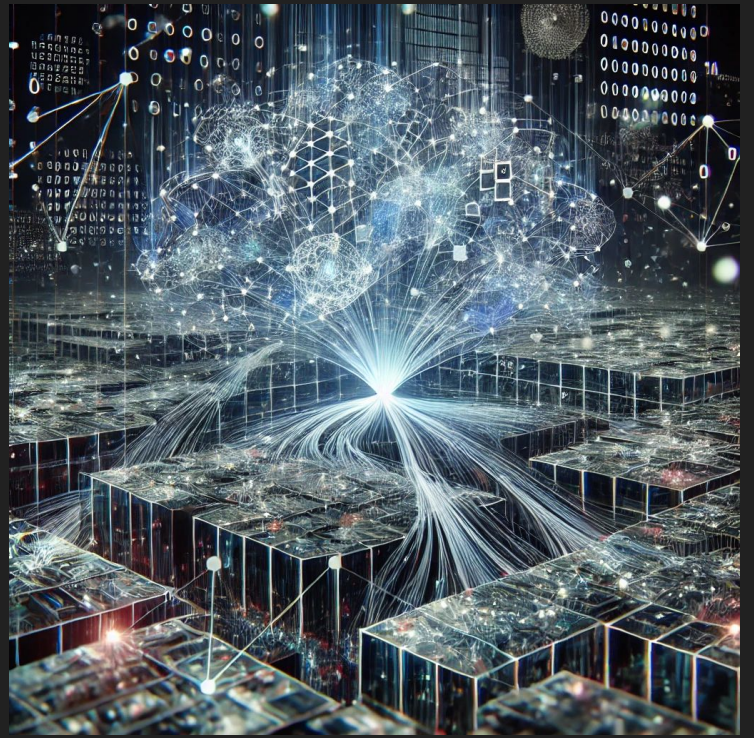
Terms & Context

ML
Machine learning in general. Pre-2022 mostly supervised learning on structured, labeled data.

Generative AI (also Large Language models)

- pre-trained on large collections of unstructured data
- learns patterns in the underlying data and can generate new instances (text/pictures/video/etc)

Remember: pre-2022 mostly older-style ML. post-2022 often is LLM/GenAI



Let’s begin:

How we got here: review previous talks and predictions.

Spoiler alert: I was wrong then and I’m probably wrong now.

Also, and for fun:

Let’s have various Gen AI chatbots critique my previous conclusions.

Previous Talk #1

All of Our ML Ideas Are Bad

(and We Should Feel Bad)

SRECON EMEA 2019

Oct, 2019

tmu@google.com

Previous Talk #1

(SRE Thing) is ML **For?**

We want ML to do the annoying/repetitive/boring (but important) stuff:

- Monitoring maintenance - updating thresholds and parameters
- Alert suppression/analysis - flagging false positives, suppressing them
- Capacity Planning/Prediction
- Validation of Canaries - stopping bad pushes
- Root Cause Analysis of Incidents/Outages
- Automatic Service/Resource Scaling

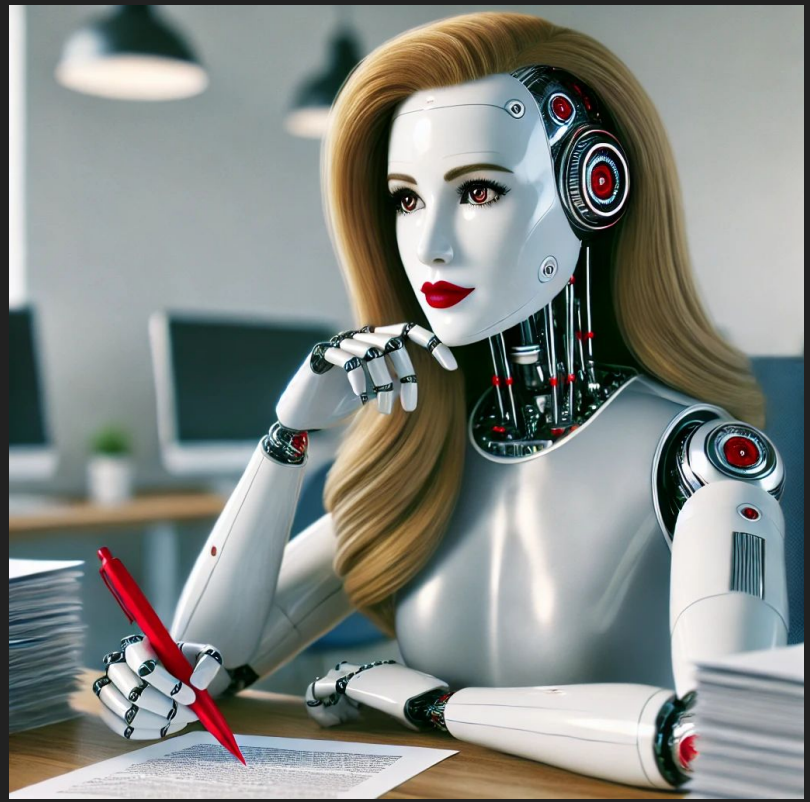
Previous Talk #1

Gen AI [*] says...

[ML] is probably most valuable right now in augmenting human operators rather than replacing them - helping with:

- *Initial problem diagnosis*
- *Documentation generation and maintenance*
- *Procedure automation planning*
- *Knowledge base search and synthesis*

Most significant: less reliance on labeled data due to un- or self-supervised pretraining.



[*] Anthropic Claude 3.5 Sonnet

Any out-of-context resemblance to a robot version of a moderately famous pop music star is obviously entirely coincidental. Haters gonna hate.

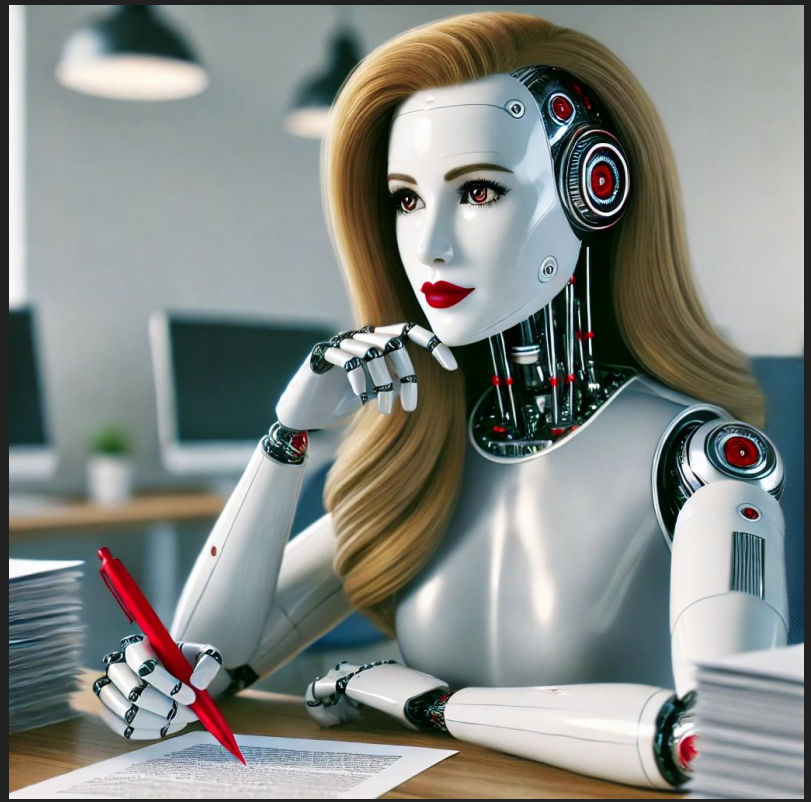
Previous Talk #1

Gen AI also says...

*"Only capacity planning and basic alert correlation had seen significant practical success [in 2019]. Most of these aspirations were still more aspirational than reality, with **production implementations being limited or experimental**." [*]*

But another said:

*"In summary, **all the mentioned use cases were plausible in 2019**, but some were more mature and widely adopted than others" [**]*



[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Images from DALL-E in ChatGPT except where otherwise noted

Previous Talk #1

Gen AI also says...

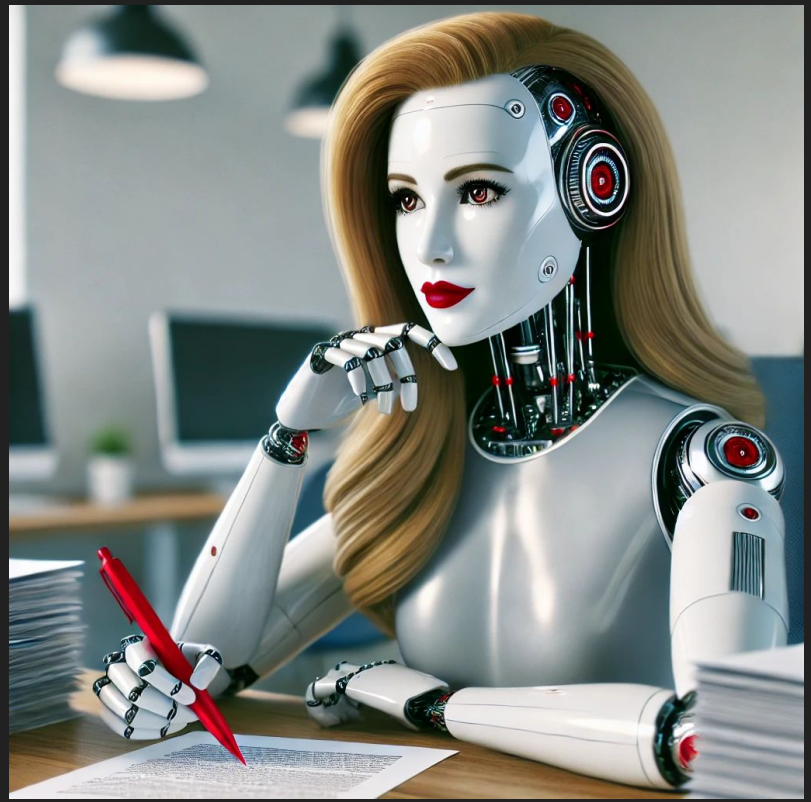
Yes

*"Only capacity planning and basic alert correlation had seen significant practical success [in 2019]. Most of these aspirations were still more aspirational than reality, with **production implementations being limited or experimental**." [*]*

But

No

*"In summary, **all the mentioned use cases were plausible in 2019**, but some were more mature and widely adopted than others" [**]*



[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Images from DALL-E in ChatGPT except where otherwise noted

Previous Talk #1

Gen AI also says (now)...

“Overall Assessment:

MOST EFFECTIVE: Capacity Planning, Automatic Scaling, Alert suppression/analysis

MODERATELY EFFECTIVE:. Monitoring maintenance, Canary validation.

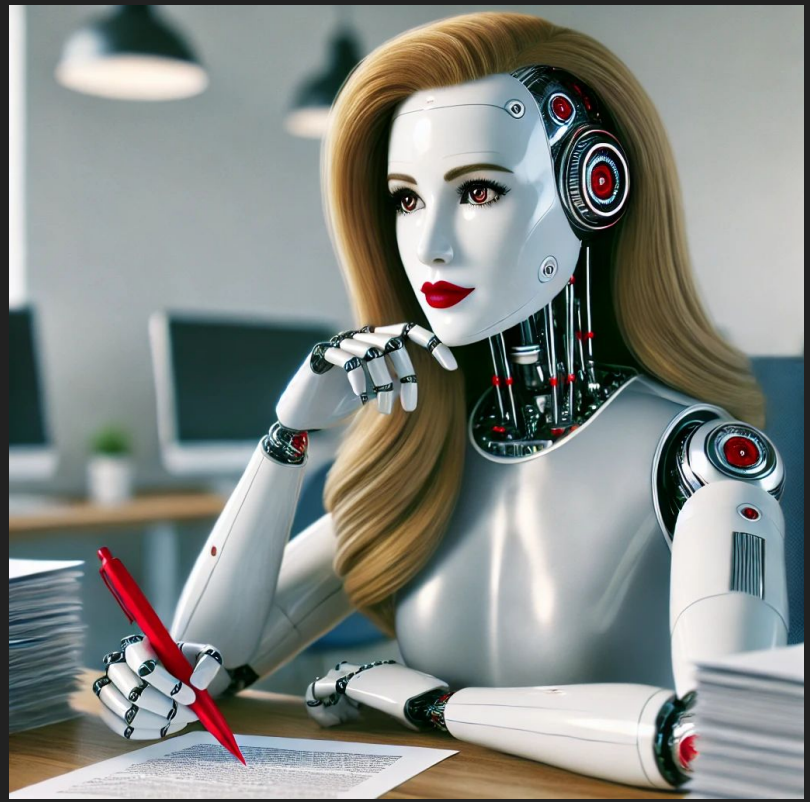
STILL CHALLENGING: Root Cause Analysis“[*]

But

“Mature, Effective Use Cases: Monitoring maintenance, alert suppression/analysis, capacity planning, automatic service/resource scaling.

Somewhat Effective, Still Evolving: Validation of canaries, root cause analysis.”[**]

[*] Anthropic Claude 3.5 Sonnet [**] Chat GPT 4o



Images from DALL-E in ChatGPT except where otherwise noted

Previous Talk #1

Gen AI also says (now)...

Yes

“MOST EFFECTIVE: Capacity Planning, Automatic Scaling, Alert suppression/analysis

MODERATELY EFFECTIVE:. Monitoring maintenance, Canary validation.

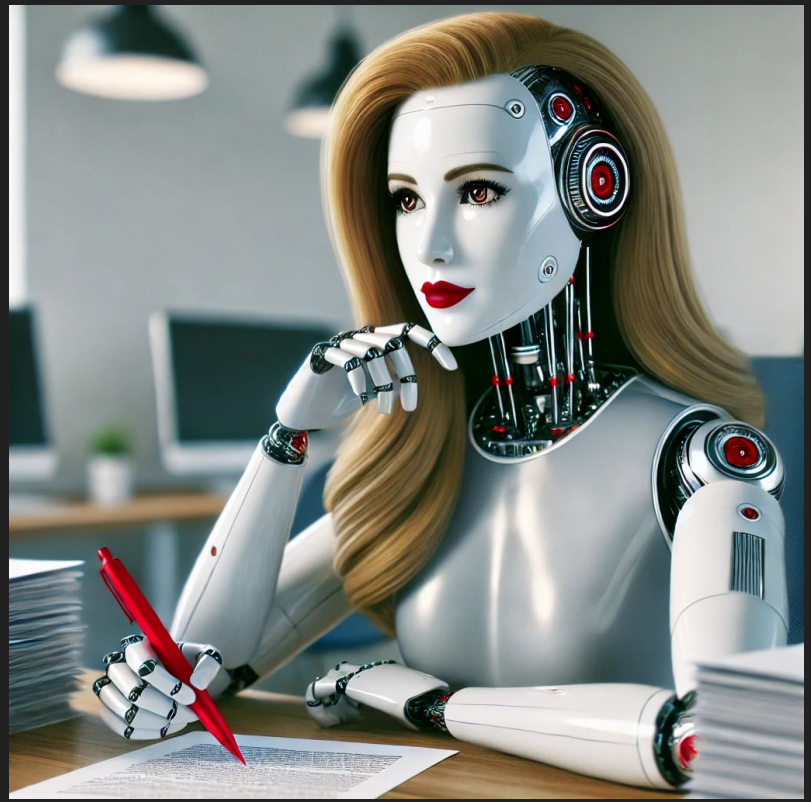
STILL CHALLENGING: Root Cause Analysis“ [*]

But

Meh

“Mature, Effective Use Cases: Monitoring maintenance, alert suppression/analysis, capacity planning, automatic service/resource scaling.

Somewhat Effective, Still Evolving: Validation of canaries, root cause analysis.”[**]



[*] Anthropic Claude 3.5 Sonnet [**] Chat GPT 4o

Images from DALL-E in ChatGPT except where otherwise noted

Previous Talk #1

Gen AI also says (now)...

Yes

“MOST EFFECTIVE: Capacity Planning, Automatic Scaling, Alert suppression/analysis

MODERATELY EFFECTIVE:. Monitoring maintenance, Canary validation.

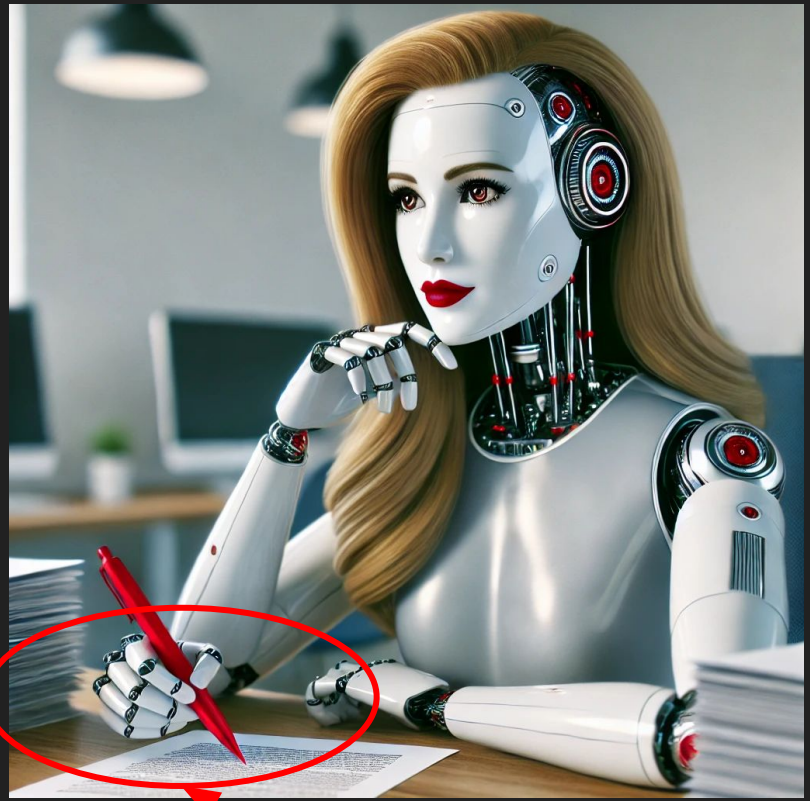
STILL CHALLENGING: Root Cause Analysis“ [*]

But

Meh

“Mature, Effective Use Cases: Monitoring maintenance, alert suppression/analysis, capacity planning, automatic service/resource scaling.

Somewhat Effective, Still Evolving: Validation of canaries, root cause analysis.”[**]



Classic AI Image

[*] Anthropic Claude 3.5 Sonnet [**] Chat GPT 4o

Images from DALL-E in ChatGPT except where otherwise noted

Previous Talk #2

Automating Operations with ML

OPML20

tmu@google.com, stross@google.com

Previous Talk #2

y Results

Problems are dynamic requiring periodic model refresh.

Human labeling is expensive compared to value for most applications.

While many ML techniques do solve Operational management problems, they often do so at a cost that is higher than the value that they provide.

One common alternative: straightforward Heuristics. These are brittle and somewhat less flexible, but are still almost as good as ML for many applications.

Of the techniques evaluated, Autoscaling is the most effective application of ML.

Photo by Skitterphoto from Pexels

Previous Talk #2

Gen AI Says, Then...

*“In 2020, the industry was experiencing a reality check about ML in operations, recognizing that **simpler solutions often provided better value**. The practical challenges of implementation and maintenance [of ML] often outweighed the benefits” [*]*

*“While ML offers powerful capabilities, its application is often weighed against the cost, with **simpler heuristics remaining a preferred approach for many use cases**. Autoscaling stands out as a strong example where ML has delivered clear, effective benefits.” [**]*

[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Previous Talk #2

Gen AI Says, Then...

Yes

*"In 2020, the industry was experiencing a reality check about ML in operations, recognizing that **simpler solutions often provided better value**. The practical challenges of implementation and maintenance [of ML] often outweighed the benefits" [*]*

Also Yes

*"While ML offers powerful capabilities, its application is often weighed against the cost, with **simpler heuristics remaining a preferred approach for many use cases**. Autoscaling stands out as a strong example where ML has delivered clear, effective benefits." [**]*

[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Previous Talk #2

Gen AI Says, Now...

"[...]While the challenges persist, the tools and techniques for addressing them have improved, making ML more practical for operational use cases than in 2020. But the need for careful consideration of value vs. complexity remains important." [*]

*While the **initial hurdles of ML in operations have been overcome**, successful implementation requires careful planning, technical expertise, and a strong understanding of business needs. [Now] organizations can harness the power of ML to drive innovation and efficiency. [**]*

[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Previous Talk #2

Gen AI Says, Now...

Yes

"[...]While the challenges persist, the tools and techniques for addressing them have improved, making ML more practical for operational use cases than in 2020. But the need for careful consideration of value vs. complexity remains important." [*]

No

*While the **initial hurdles of ML in operations have been overcome**, successful implementation requires careful planning, technical expertise, and a strong understanding of business needs. [Now] organizations can harness the power of ML to drive innovation and efficiency. [**]*

[*] Anthropic Claude 3.5 Sonnet
[**] Chat GaPT 4o

Previous Talk #2

Gen AI Says, Now...

Yes

"[...]While the challenges persist, the tools and techniques for addressing them have improved, making ML more practical for operational use cases than in 2020. But the need for careful consideration of value vs. complexity remains important." [*]

No

*While the **initial hurdles of ML in operations have been overcome**, successful implementation requires careful planning, technical expertise, and a strong understanding of business needs. [Now] organizations can harness the power of ML to drive innovation and efficiency. [**]*

Content Warning: Relentless AI positivity common in LLM output

[*] Anthropic Claude 3.5 Sonnet
[**] Chat GPT 4o

Previous Paper

for Operations

Pitfalls, Dead Ends, and Hope

STEVEN ROSS AND TODD UNDERWOOD

Steven Ross is a Technical Lead in site reliability engineering for Google in Pittsburgh, and has worked on machine learning at Google since Pittsburgh Pattern Recognition was acquired by Google in 2011. Before that he worked as a Software Engineer for Dart Communications, Fishtail Design Automation, and then Pittsburgh Pattern Recognition until Google acquired it. Steven has a BS from Carnegie Mellon University (1999) and an MS in electrical and computer engineering from Northwestern University (2000). He is interested in mass-producing machine learning models. stross@google.com

Todd Underwood is a lead Machine Learning for Site Reliability Engineering Director at Google and is a Site Lead for Google's Pittsburgh office. ML SRE teams build and scale internal and external ML services and are critical to almost every product area at Google. Todd was in charge of operations, security, and peering for Renesys's Internet intelligence services that is now part of Oracle's cloud service. He also did research for some early social products that Renesys worked on. Before that Todd was Chief Technology Officer of Oso Grande, an independent Internet service provider (AS2901) in New Mexico. Todd has a BA in philosophy from Columbia University and a MS in computer science from the University of New Mexico. He is interested in how to make computers and people work much, much better together. tmu@goggle.com

Machine learning (ML) is often proposed as the solution to automate this unpleasant work. Many believe that ML will provide near-magical solutions to these problems. This article is for developers and systems engineers with production responsibilities who are lured by the siren song of magical operations that ML seems to sing. Assuming no prior detailed expertise in ML, we provide an overview of how ML works and doesn't, production considerations with using it, and an assessment of considerations for using ML to solve various operations problems.

Even in an age of cloud services, maintaining applications in production is full of hard and tedious work. This is unrewarding labor, or toil, that we collectively would like to automate. The worst of this toil is manual, repetitive, tactical, devoid of enduring value, and scales linearly as a service grows. Think of work such as manually building/testing/deploying binaries, configuring memory limits, and responding to false-positive pages. This toil takes time from activities that are more interesting and produce more enduring value, but it exists because it takes just enough human judgment that it is difficult to find simple, workable heuristics to replace those humans.

We will list a number of ideas that appear plausible but, in fact, are not workable.

What Is ML?

Machine learning is the study of algorithms that learn from data. More specifically, ML is the study of algorithms that enable computer systems to solve some specific problem or perform some task by learning from known examples of data. Using ML requires training a model on data where each element in the data has variables of interest (features) specified for it. This training creates a model that can later be used to make inferences about new data. The generated model is a mathematical function, which determines the predicted value(s) ("dependent variable(s)") based on some input values ("independent variables"). How well the model's inferences fit the historical data is the objective function, generally a function of the difference between predictions and correct inferences for supervised models. In an iterative algorithm, the model parameters are adjusted incrementally on every iteration such that they (hopefully) decrease the objective function.

Main Types of ML

In order to understand how we'll apply ML, it is useful to understand the main types of ML and how they are generally used. Here are broad categories:

?

!

!

Google

;login: issue: Summer 2020, Vol. 45, No. 2

Previous Paper

Conclusion

While ML is promising for many applications, it is difficult to apply to operations today because it makes errors, it requires a large amount of high-quality training data that is hard to obtain and label correctly, and it's hard to explain the reasons behind its decisions. We've identified some areas where people commonly think ML can help in operations and what makes it difficult to use in those applications. We recommend using standard tools to improve operations first before moving forward with ML, and we suggest collecting and curating your training data as the first step to take before using ML in operations.

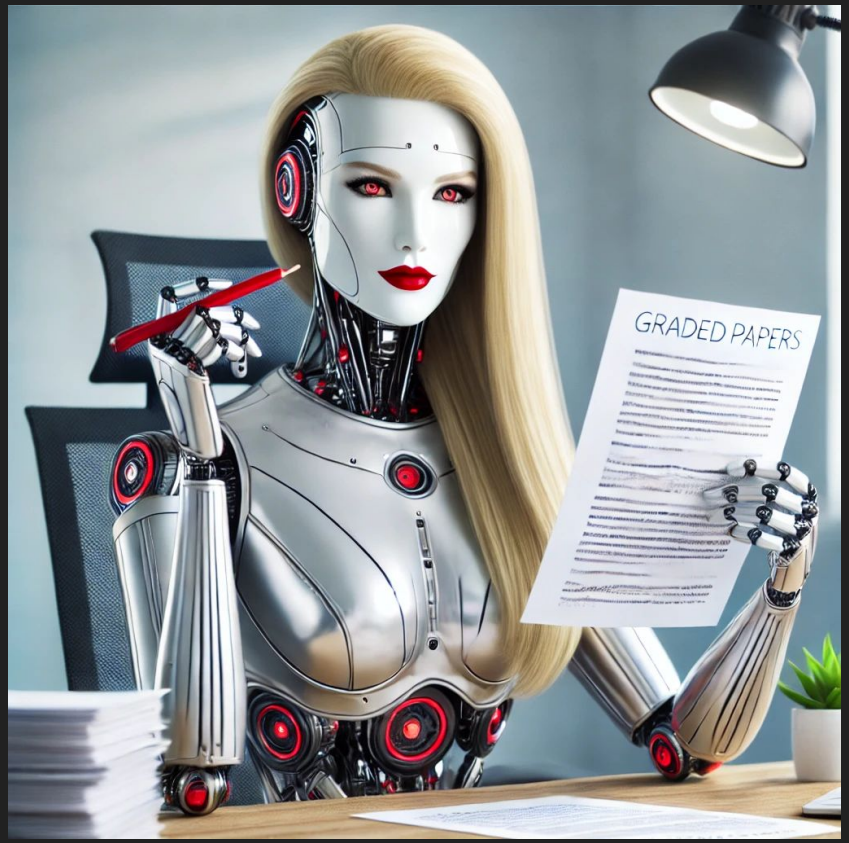
Google

Previous Paper

Gen AI says...

“The overall statement was quite prescient and aligned with practical experience in 2020. Many organizations were discovering that ML wasn't a silver bullet for operational challenges, and that fundamental operational excellence was a prerequisite for successful ML implementation.” [*]

“The statement in 2020 is partially true. While it accurately reflects some of the challenges associated with applying ML to operations at that time, it also underestimates the potential and progress being made in the field.” [**]



[*] Anthropic Claude 3.5 Sonnet
[**] Google Gemini

Previous Paper

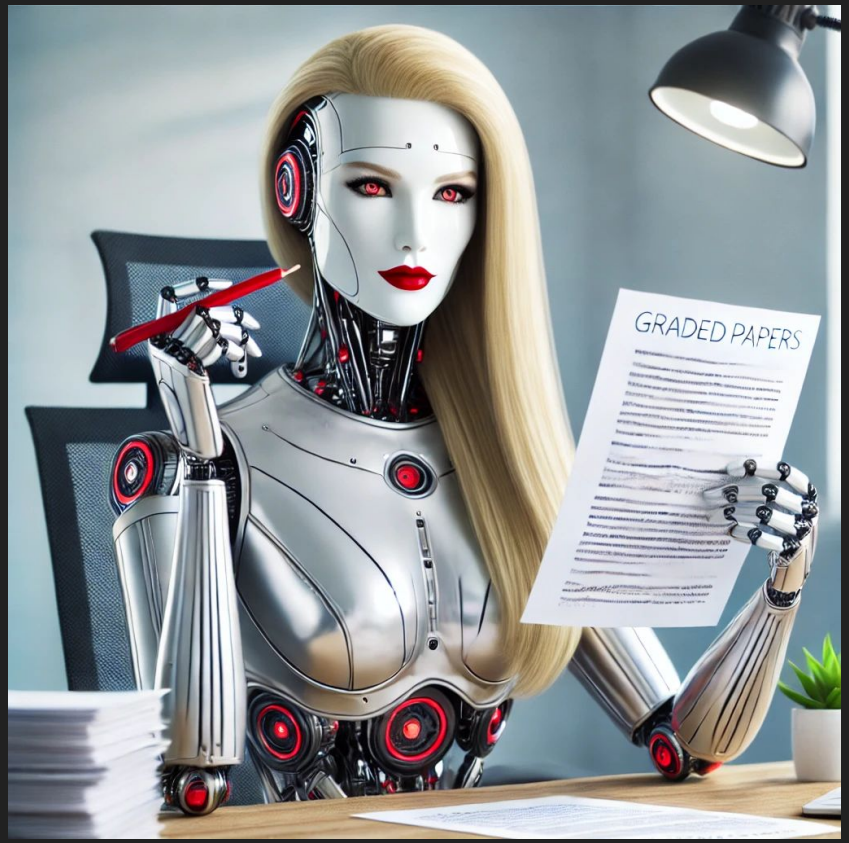
Gen AI says...

Yes

“The overall statement was quite prescient and aligned with practical experience in 2020. Many organizations were discovering that ML wasn't a silver bullet for operational challenges, and that fundamental operational excellence was a prerequisite for successful ML implementation.” [*]

Meh

“The statement in 2020 is partially true. While it accurately reflects some of the challenges associated with applying ML to operations at that time, it also underestimates the potential and progress being made in the field.” [**]



[*] Anthropic Claude 3.5 Sonnet
[**] Google Gemini

Previous Paper

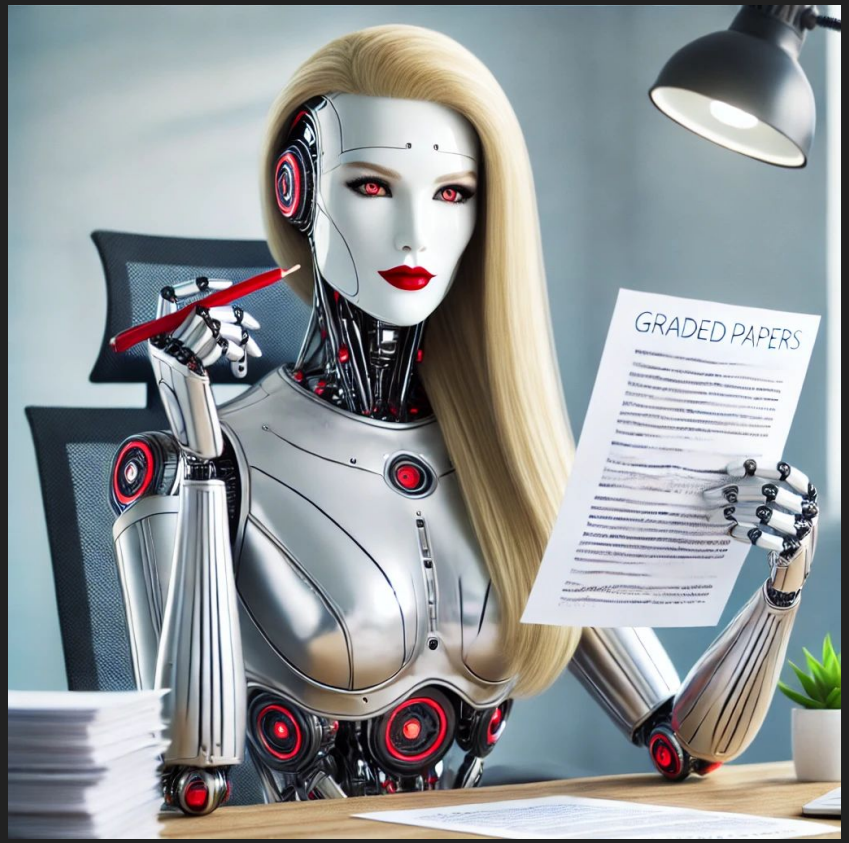
Gen AI says...

Yes

“The overall statement was quite prescient and aligned with practical experience in 2020. Many organizations were discovering that ML wasn't a silver bullet for operational challenges, and that fundamental operational excellence was a prerequisite for successful ML implementation.” [*]

Meh

“The statement in 2020 is partially true. While it accurately reflects some of the challenges associated with applying ML to operations at that time, it also underestimates the potential and progress being made in the field.” [**]



Content Warning: Relentless AI positivity common in LLM output

[*] Anthropic Claude 3.5 Sonnet
[**] Google Gemini

Previous Talk #3

SRE for ML

The First 10 Years and the Next 10

Todd Underwood @tmu ♦
tmu@google.com ♦
2021-Oct ♦
SRECon ML Track

Previous Talk #3

Make ML Boring

Five Big Things Will Happen

1. Organizations will accept “older” ML technology, and AutoML approaches. Good enough will be good enough.
2. Platforms will converge in features and stabilize. And mostly fade into the background
3. Training and serving costs will plummet.
4. APIs to integrate ML into applications will stabilize and become ubiquitous
5. ML model quality evaluation will become ubiquitous and trustworthy

Previous Talk #3

Make ML Boring

Five Big Things Will Happen

1. Organizations will accept “older” ML technology, and AutoML approaches. Good enough will be good enough. **[Yes, but… GenAI happened]**
2. Platforms will converge in features and stabilize. And mostly fade into the background. **[Not really]**
3. Training and serving costs will plummet. **[No!]**
4. APIs to integrate ML into applications will stabilize and become ubiquitous. **[Somewhat true.]**
5. ML model quality evaluation will become ubiquitous and trustworthy. **[Not really]**

Previous Talk #3

Make ML Boring

Please see me after class

F

Five Big Things Will Happen

1. [illegible] will accept "older" ML technology, a[illegible] AutoML approaches. Good enough will be [illegible]od enough. **[Yes, but… GenAI happened]**
2. [illegible]latforms will converge in features and stabilize. And mostly fade into the background. **[Not really]**
3. Training and serving costs will plummet. **[No!]**
4. APIs to integrate ML into applications will stabilize and become ubiquitous. **[Somewhat true.]**
5. ML model quality evaluation will become ubiquitous and trustworthy. **[Not really]**

Previous Talk
#4

SRE and ML:

~~Why~~ Does It Matter~~s~~*?

[*] Yet

todd underwood — Google — tmu@google.com
@tmu on twitter
linkedin.com/in/toddunder
sre.google

Previous Talk #4

e Path Ahead:

Predictions without Evidence

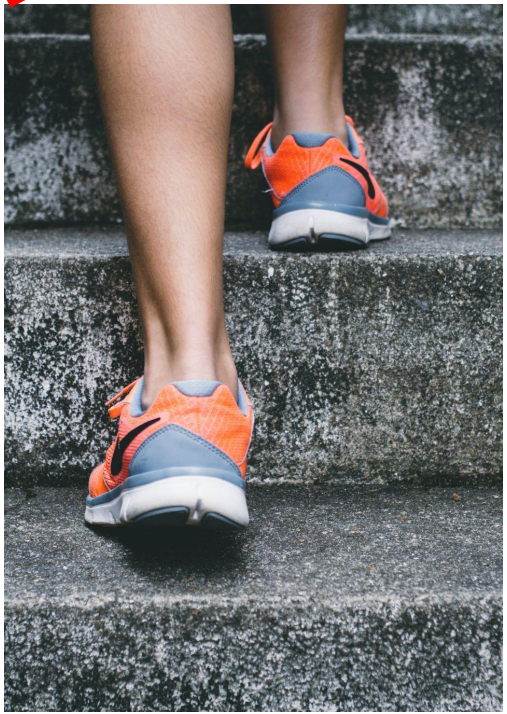


Photo by Bruno Nascimento on Unsplash

Career changes

SREs will have to do ML (**ML Reliability Engineers**?)

but we will also need **data reliability engineers**

and model quality will have to be baked in to production engineering responsibilities. (Quality as the end-to-end SLO)

Ethics

(Wildcard opinion!)

AI/ML raises huge and complicated ethical issues. SREs will be again (as we usually have been) at the forefront of addressing those issues as models enter into the world

Previous Talk #4

e Path Ahead:

Predictions without Evidence

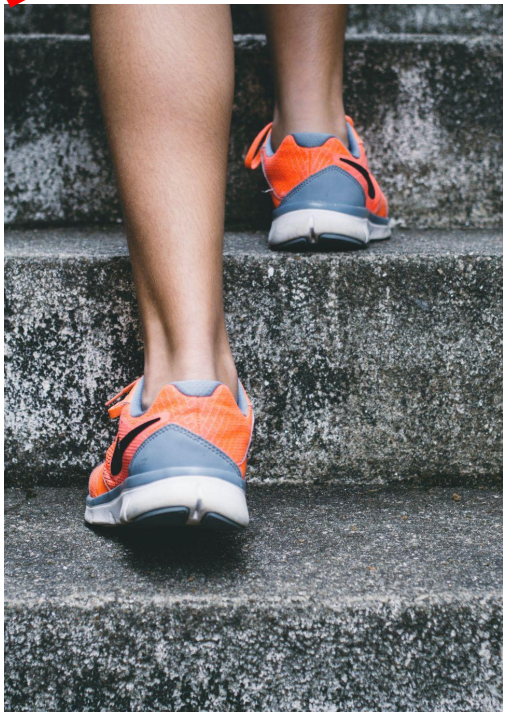


Photo by Bruno Nascimento on Unsplash

~ Yes

Career changes

SREs will have to do ML (**ML Reliability Engineers**?)

but we will also need **data reliability engineers**

and model quality will have to be baked in to production engineering responsibilities. (Quality as the end-to-end SLO)

Ethics

(Wildcard opinion!)

AI/ML raises huge and complicated ethical issues. SREs will be again (as we usually have been) at the forefront of addressing those issues as models enter into the world

Not Yet

Overall Grade:
[in the optimistic tone of an LLM]
Almost passing!

GenAI/LLMs Now:
Are they good at anything (yet)?

You can’t believe a word I say, most days, but this time it’s true.

AI Hype

We have been promised many things.

We have been sold many things.

More specifically: We have been told that the AI will run the computers instead of us.

Let’s specifically talk about AIOps.

AIOps hype

servicenow

Make self-healing IT infrastructure with AI-powered service operations

- Anomaly detection
- Outlier detection
- Malware traffic detection
- Vulnerability detection
- Historical analysis
- Performance analysis
- Root cause analysis
- Remediation advice

bmc

Intro

AIOps helps you analyze data, identify patterns, predict problems, and resolve issues faster, making your entire business more resilient.

7 Benefits of AIOps

- Improved Time Management and Prioritization
- IT Spend Reduction
- Accelerated Innovation
- More Collaboration
- Automation at Scale
- Digital Transformation

AIOps shares overlapping with SRE. It uses business operations' massive data and ML-sourced predictive insights to help site reliability engineers reduce incident resolution time.

Get to the precise root cause and eliminate alert storms

What are the benefits of AIOps?

Network Operations Center (NOC) staff, IT Operations teams, DevOps engineers, and site reliability engineers (SREs) all benefit from AIOps in the following situations:

1. **Proactively address issues before they impact performance and cause downtime:** Complex IT

AIOps hype

servicenow

Make self-healing IT infrastructure with AI-powered service operations

- Anomaly detection
- Outlier detection
- Malware traffic detection
- Vulnerability detection
- Historical analysis
- Performance analysis

bmc

The Intelligence Age

September 23, 2024

site reliability engineers reduce incident resolution time.

teams, DevOps engineers, and site reliability engineers (SREs) all benefit from AIOps in the following situations:

1. **Proactively address issues before they impact performance and cause downtime:** Complex IT

Get to the precise root cause and eliminate alert storms

(Occasionally acknowledged) Challenges

- **Time to value**: AIOps systems can be difficult to design, implement, deploy and manage, so it can take some time to see any return on investment
- **Data**: the volume, quality and consistency of data produced by modern IT operations can be overwhelming and difficult to wrangle into something that can be used for modelling

Paying double

management, studies show that only about half (53%) of AI projects accomplish the move from prototype to production. **Several factors add to that difficu**

Missing the big picture

AIOps is great if you need to remediate relatively simple and straightforward problems that arise in the context of monitoring and management.

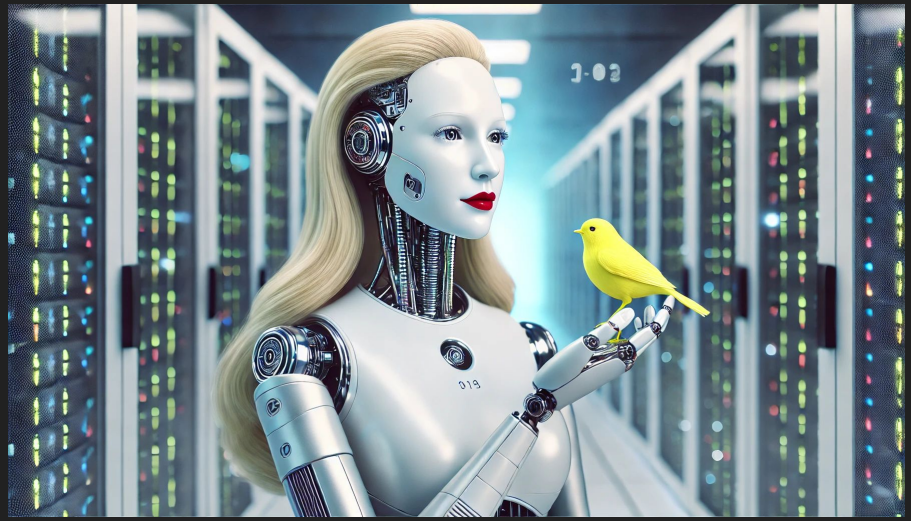
These are the tears of AI hype ricocheting off of reality.

What LLMs are Bad at (according to LLMs)

The greatest applications of LLM of all time were never made.

These applications for AI don't work:

- Root Cause Analysis
- Canary Validation
- Paying off (being cost efficient)



Is there anything LLMs are good at?

This is going to be boring.

Some small use cases that are meh:

- coding
  - genuinely useful
  - getting better
  - requires substantial expertise
- bureaucracy
  - performance reviews
  - job descriptions
  - some documentation

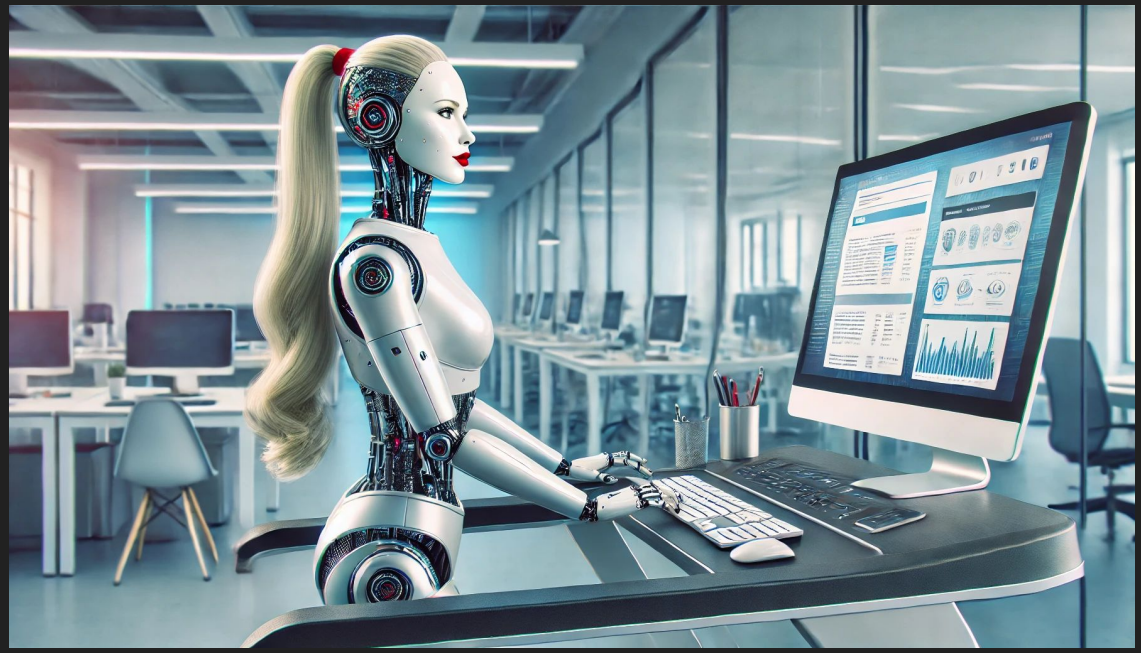
What is the pen for? Is she going to write on the monitor?

Key use case now: Knowledge base / search

Train a model on everything everyone needs to know:

- All documents (google docs, notion, word, wiki)
- All slack/teams/irc/chat logs
- All code

Create a straightforward url (go/what or go/ask or whatever)



Why?

Super fast onboarding.

Fast troubleshooting and education.

Internal communications++ (find the right person/team)

In short:

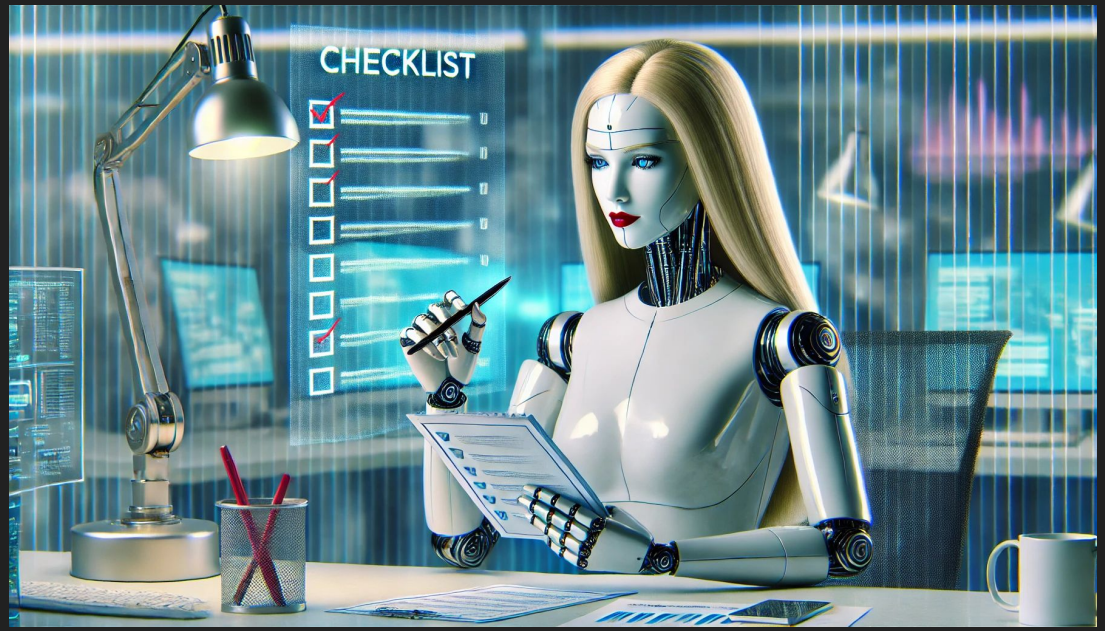
It works.

Anything Else?

Document summarization (but has to be checked).

Code and config auditing (but writing tests would be better).

Monitoring tuning (but needs to be checked).



Parting thoughts: A Prediction

I think I've seen this film before. I didn't like the ending.
But this is me trying.

Agents will finally happen

Now:

- Anthropic agents
- Google rumo(u)red to have something in flight
- Startup arcade-ai.com full internet API in early release.

Why?

- flexible APIs to other data/services will rapidly identify best use cases for AI

When?

- Soon? We don’t want to waste all of this AI potential.

In Conclusion

Previous predictions were ~meh

LLMs are not yet great at most stuff.

They’re super good at one thing.

Might get better with Agents.

Thank you.

I had a marvelous time ruining

(some of) the AI hype.